EXPLORATORY MODEL SELECTION - SOME EXAMPLES

BU-1566 -M **May 2001**

Walter T. Federer

Keywords: Orthogonal polynomial regression, differential gradients, replication, interaction.

Abstract:

Exploratory model selection was used to find a response model that accounted for the spatial variability present in the experimental results from four examples of field experiments. it was found that the class of differential gradients within incomplete blocks was useful for finding a response model that accounted for the spatial variability present in the first example. The class of orthogonal polynomial regression of response on row and column position and interactions of the regressions was useful for discovering an appropriate response model for the data of examples two, three, and four. The results obtained from the selected response model were compared with standard textbook analyses. Considerable differences in residual mean squares, coefficients of variation, and F-values for treatment to residual mean squares were found. The increase in replication for the selected response model over textbook response models is demonstrated. The increase can be many fold.

EXPLORATORY MODEL SELECTION – SOME EXAMPLES

by

Walter T. Federer

ABSTRACT

Exploratory model selection was used to find a response model that accounted for the spatial variability present in the experimental results from four examples of field experiments. It was found that the class of differential gradients within incomplete blocks was useful for finding a response model that accounted for the spatial variability present in the first example. The class of orthogonal polynomial regression of response on row and column position and interactions of the regressions was useful for discovering an appropriate response model for the data of examples two, three, and four. The results obtained from the selected response model were compared with standard textbook analyses. Considerable differences in residual mean squares, coefficients of variation, and F-values for treatment to residual mean squares were found. The increase in replication for the selected response model over textbook response models is demonstrated. The increase can be many fold.

Key words: Orthogonal polynomial regression, differential gradients, replication, interaction, functions of row and column effects, fixed effect analysis, random effect analysis, computer code.

BU-1566-M May 2001

In the Technical Report Series of the Department of Biometrics, Cornell University, Ithaca, New York 14853

EXPLORATORY MODEL SELECTION – SOME EXAMPLES

by

Walter T. Federer

INTRODUCTION

Data from four different examples involving field experiments were examined to determine an appropriate response model that accounted for the variability present in the experimental results. An experimenter selects an experiment design plan that is thought appropriate for a forthcoming field experiment. Then the experiment is laid out in a field. It should be noted that the selected experiment design *and* the actual field layout determine the design for an experiment as far as spatial variation in an experiment is concerned. Also, events that occur during the conduct of an experiment need to be taken into account when analyzing the data. The direction of the spatial variation may not coincide with blocking pattern used for the experiment. Several types of events can occur during the course of an experiment that determines the pattern of variation. Since an experimenter is not blessed with the knowledge of the spatial variation patterns in an experiment, it is necessary to attempt to find a response model that accounts for the variability present in an experiment. In exploratory model selection, the data analyst sets up a class of plausible response models and then selects one that best accounts for the spatial variation present in the experiment. This needs to be done for each characteristic measured in an experiment as a different response model may be appropriate for each one.

The first example is an incomplete block experiment design arrangement with the incomplete blocks laid one below the other to form a row-column layout within each complete block. Examples two and three were laid out in a 15-row by 12-column arrangement. The fourth example was designed as a randomized complete block experiment design but laid out in an 8-row by 7-column arrangement. In the first example two classes of response models were examined for the character weight of grain. These were differential gradients within incomplete blocks and orthogonal polynomial regression of response on row and column position and interaction of row and column regressions. Since the latter did not account for the variation in grain weight, the differential regression method was used for the other six characters measured in this experiment. The class of orthogonal polynomial regressions for rows, columns, and interactions was appropriate for finding a response model for the data of examples two, three, and four.

INCOMPLETE BLOCK EXPERIMENT DESIGN EXAMPLE

An incomplete block design experiment design example was obtained from Dr. Matthew Reynolds, International Center for Maize and Wheat Improvement (CIMMYT). There were v = 120 wheat genotypes arranged in b = 15 incomplete blocks of size k = 8 in each of r = 2 replicates. Seven different responses were obtained for each of the wheat genotypes. They were weight of grain (grainwt), rust infection index, grain weight per meter squared (grm2), maturity, anthesis, total green weight (TGW), and CTD X. Two types of response models for accounting for the spatial variation present in the experiment were examined. These were trend analyses using orthogonal polynomial regression coefficients for response on position in row or column and interactions of these regressions (regression method) and differential orthogonal polynomial regressions on position within the incomplete blocks (differential gradients). The first method was used only for weight of grain owing to the fact that the differential gradient method was so much more effective in accounting for the spatial variation present in the experiment.

For the response weight of grain per plot (experimental unit), the standard textbook analysis for an incomplete block design (IBD) resulted in a residual (error) mean square of 52,395 and a coefficient of variation of 10.9%. Since the experiment was laid out in a 15-row (block) by 8-column arrangement within each complete block (replicate), orthogonal polynomial regression coefficients of grain weight on position in row and in column were obtained. Also, the interactions of all row and column regressions were examined. The Bozivich, Bancroft , and Hartley (1956) rule used by Federer, Crossa, and Franco (1998) resulted in the following response model

Grainwt = mean + C2 + R1 + R3 + R6 + R10 + R12 +C1*R12 + C1*R14 + C2*R2 + C2*R6 + C2* R10 + C3*R3 + C3*R4 + C3*R5 + C3*R7 + C3*R13 + C4* R4 + C4*R6 + C4*R7 + C4*R8 + C5*R1 + C5*R4 + C5*R5 + C5*R8 + C5*R9 + C6*R10 + C6*R13 + C6*R14 + C7*R6 + C7*R7 + C7*R13 + error

where Ri is the ith orthogonal polynomial regression coefficient of grain weight on row position and Cj is the jth orthogonal polynomial regression coefficient of grain weight on column position. The asterisk denotes an interaction as used in SAS. The residual mean square for the above response model was 54,678 with 84 degrees of freedom and a coefficient of variation of 11.2%. Thus the best regression model using the selection rule resulted in a residual mean that was larger than the textbook analysis for an incomplete block experiment design.

The differential regression method was effective in accounting for more of the spatial variation than the above as shown below. Using the response model for weight of grain

grainwt = mean + replicate + genotype + block(replicate) + Cj*block(replicate) + error,

the following results (with the minimum listed in bold type) were obtained

C1	C2	C3	C4	C5	C6	C7
41,604	50,477	53,058	65,257	54,811	52,640	48,207
9.7%	10.7%	11.0%	12.2%	11.2%	11.0%	10.5%

where the first row is Cj, the second row is residual mean square, and the third row is the coefficient of variation. Adding the term C1*block(replicate) to the incomplete block response model was effective in accounting for a sizeable portion of the residual variation. These residual mean squares are associated with 61 degrees of freedom. Since there are sufficient degrees of freedom to search further, a second term of Ch*block(replicate) for h not equal to j, is added to the above response model equation . All possible pairs were investigated. The residual mean square results with the associated coefficient of variation obtained are (the minimum listed in bold type)

	C2	C3	C4	C5	C6	C7
C1	42,504	38,645	54,444	36,558	43,316	**31,546**
	9.8%	9.4%	11.1%	9.1%	9.9%	**8.5%**
C2	--	52,368	69,244	56,841	42,918	43,292
	--	10.9%	12.6%	11.4%	9.9%	9.9%
C3	--	--	72,965	61,546	55,971	43,834
	--	--	12.9%	11.8%	11.3%	10.0%
C4	--	--	--	72,202	71,772	68,543
	--	--	--	12.8%	12.8%	12.5%
C5	--	--	--	--	57,240	58,136
	--	--	--	--	11.4%	11.5%
C6	--	--	--	--	--	46,811
	--	--	--	--	--	10.3%

Adding the term C7*block(replicate) to the response model resulted in a residual mean square of 31,546 (listed in bold type) and a coefficient of variation of 8.5%. The ratio of this mean square to the IBD mean square is 31,546/52,395 = 0.602, which is 40% smaller than the standard textbook incomplete block analysis residual mean square. The addition of this term over the C1*block((replicate) term is sizeable, i.e., 31,546/41,604 = 0.758, or a 24% reduction. Using the textbook analysis for the incomplete block design (IBD) would require 52,395/31,546 = 1.66 times more replication than the differential gradient model with two terms in the response model. Also, even with allocating an additional 30 degrees of freedom to spatial variation, there are still 31 degrees of freedom associated with the residual mean square.

For the response rust index, the IBD response model gave a residual mean square of 0.239664 and a coefficient of variation of 24.6%. Adding Cj*block(replicate) to the IBD response model resulted in the following mean squares and coefficients of variation

C1	C2	C3	C4	C5	C6	C7
0.181704	0.227638	0.199817	0.310642	0.241341	0.257496	0.268095
21.4%	24.0%	22.4%	28.0%	24.7%	25.5%	26.0%

Adding all possible pairs of the Cj*block(replicate) terms to the IBD response model for rust index resulted in the following residual mean squares and coefficients of variation:

	C2	C3	C4	C5	C6	C7
C1	0.139226	**0.128184**	0.224499	0.173218	0.275335	0.200575
	18.7%	**18.0%**	23.8%	20.9%	26.3%	22.5%
C2	--	0.193638	0.325319	0.205132	0.151301	0.229842
	--	22.1%	28.6%	22.7%	19.5%	24.1%
C3	--	--	0.258216	0.212639	0.172752	0.132256
	--	--	25.5%	23.2%	20.9%	18.3%
C4	--	--	--	0.367959	0.327632	0.383353
	--	--	--	30.5%	28.7%	31.1%
C5	--	--	--	--	0.220008	0.264390
	--	--	--	--	23.6%	25.8%
C6	--	--	--	--	--	0.315708
	--	--	--	--	--	28.2%

Adding the two terms C1*block(replicate) and C3*block(replicate) to the IBD response model resulted in a 100(1 - 0.128184/0.239664) = 46.5% decrease in the residual mean square. The additional term of C3*block(replicate) resulted in a decrease of 100(1 – 0.128184/0.181704) = 29.5%. Using the two terms C3*block(replicate) and C7*block(replicate) to the IBD response model resulted in approximately the same residual mean square.

The IBD response model for total green weight, TGW, was 2.63540 and the coefficient of variation was 6.5%. Adding Cj*block(replicate to the IBD response model resulted in the following:

C1	C2	C3	C4	C5	C6	C7
1.99125	2.30948	2.83428	3.03865	2.68701	2.36558	2.48401
5.6%	6.1%	6.7%	7.0%	6.5%	6.1%	6.3%

Adding all possible pairs of Cj*block(replicate) to the IBD response model resulted in the following residual mean squares and coefficients of variation:

	C2	C3	C4	C5	C6	C7
C1	1.99019	2.11613	2.08429	1.78563	1.65424	2.39457
	5.6%	5.8%	5.8%	5.3%	5.1%	6.2%
C2	--	2.27947	2.86999	2.27355	2.15619	**1.55874**
	--	6.0%	6.8%	6.0%	5.9%	**5.0%**
C3	--	--	3.78479	2.63581	2.31121	2.41137
	--	--	7.8%	6.5%	6.1%	6.2%
C4	--	--	--	2.79232	3.01475	3.09224
	--	--	--	6.7%	6.9%	7.0%
C5	--	--	--	--	2.31945	2.20337
	--	--	--	--	6.1%	5.9%
C6	--	--	--	--	--	2.07578

--	--	--	--	--	5.8%

Adding the term C1*block(replicate) reduced the residual mean square from 2.63540 for the IBD response model to 1.99125 for a 24.4% reduction. Adding the terms C2*block(replicate) and C7*block(replicate) to the IBD model reduced the residual mean square to 1.55874 for a 40.9% reduction over that for the IBD residual mean square.

The residual mean square for grain weight per meter squared, grm2, using the IBD response model was 802,741 and the coefficient of variation was 10.8%. Adding the term Cj*block(replicate) to this response model resulted in the following residual mean squares and coefficients of variation:

C1	C2	C3	C4	C5	C6	C7
666,921	719,217	865,894	960,822	848,408	834,343	**652,584**
9.8%	10.2%	11.2%	11.8%	11.1%	11.0%	**9.7%**

Adding all possible pairs of the terms Cj*block(replicate) to the IBD response model resulted in the following residual mean squares and coefficients of variation:

	C2	C3	C4	C5	C6	C7
C1	670,560	548,440	828,053	796,350	733,292	**518,515**
	9.8%	8.9%	10.9%	10.7%	10.3%	**8.7%**
C2	--	787,692	900,923	855,114	661,144	555,981
	--	10.7%	11.4%	11.1%	9.8%	9.0%
C3	--	--	1,049,641	985,465	1,066,514	719,430
	--	--	12.3%	11.9%	12.4%	10.2%
C4	--	--	--	1,088,518	1,040,759	865,512
	--	--	--	12.5%	12.3%	11.2%
C5	--	--	--	--	813,219	665,287
	--	--	--	--	10.8%	9.8%
C6	--	--	--	--	--	691,924
	--	--	--	--	--	10.0%

Adding the two terms C1*block(replicate) and C7*block(replicate) to the IBD response model for grm2 resulted in a decrease in the residual mean square of 100(1 – 518,515/802,741) = 35.4%. The term C7*block(replicate) resulted in a decrease in the residual mean square of 100(1 – 518,515/666,921) = 22.3%.

For the character anthesis, the IBD response model resulted in a residual mean square of 10.55307 and a coefficient of variation of 5.6%. When the term Cj*block(replicate) was added to the IBD response model, the following residual mean squares and coefficients of variation were obtained:

C1	C2	C3	C4	C5	C6	C7
2.74642	11.76045	11.25254	11.39026	11.95179	13.41503	11.95456
2.8%	5.9%	5.8%	5.8%	5.9%	6.3%	5.9%

Adding all possible pairs to the IBD response model resulted in the following residual mean squares and coefficients of variation:

	C2	C3	C4	C5	C6	C7
C1	**1.39405**	1.81953	2.14683	2.13473	2.97206	2.27563
	2.0%	2.3%	2.5%	2.5%	3.0%	2.6%
C2	--	13.41809	12.62190	14.77821	17.37477	13.66489
	--	6.3%	6.1%	6.6%	7.2%	6.3%
C3	--	--	9.07679	10.08458	16.22906	13.29218
	--	--	5.2%	5.4%	6.9%	6.3%
C4	--	--	--	15.16121	16.24249	10.71964
	--	--	--	6.7%	6.9%	5.6%
C5	--	--	--	--	14.90119	14.95308
	--	--	--	--	6.6%	6.6%
C6	--	--	--	--	--	16.24790
	--	--	--	--	--	6.9%

Despite the small coefficient of variation, 5.6%, the addition of the two terms C1*block(replicate) and C2*block(replicate) to the IBD response model resulted in a decrease in the residual mean square of 100(1 – 1.39405/10.55307) = 86.8%. The addition of C2*block(replicate) resulted in a decrease of 100(1 – 1.39405/2.74642) = 49.2% in the residual mean square.

The IBD response model for the character maturity resulted in a residual mean square of 0.746855 and a coefficient of variation of 4.2%. Adding the term Cj*block(replicate) to the IBD response model resulted in the following residual mean squares and coefficients of variation:

C1	C2	C3	C4	C5	C6	C7
0.776706	**0.642376**	0.759444	0.826918	0.845985	0.684042	0.764738
4.3%	**3.9%**	4.2%	4.4%	4.4%	4.0%	4.2%

Adding all possible pairs of Cj*block(replicate) to the IBD response model resulted in the following:

	C2	C3	C4	C5	C6	C7
C1	0.661020	0.785610	0.834726	0.985409	0.673051	0.685046
	3.9%	4.3%	4.4%	4.8%	4.0%	4.0%
C2	--	0.661024	0.611964	0.735200	0.623386	0.623112
	--	3.9%	3.8%	4.1%	3.8%	3.8%
C3	--	--	0.990799	0.871645	**0.607233**	0.665848
	--	--	4.8%	4.5%	**3.8%**	3.9%
C4	--	--	--	1.089253	0.952376	0.862107
	--	--	--	5.0%	4.7%	4.5%
C5	--	--	--	--	0.720740	0.828290
	--	--	--	--	4.1%	4.4%
C6	--	--	--	--	--	0.683112
	--	--	--	--	--	4.0%

Adding the term C2*block(replicate) to the IBD response model reduced the residual mean square to 0.642376 from 0.746855, or a reduction of 14.0%. Adding the two terms C3*block(replicate) and C6*block(replicate) further reduced the residual mean square to 0.607233, or a reduction of 18.7% over that obtained for the IBD response model.

The IBD response model for the character CTD X resulted in a residual mean square of 0.146648 and a coefficient of variation of 13.8%. Adding Cj*block(replicate) to the IBD response model resulted in the following residual mean squares and coefficients of variation:

C1	C2	C3	C4	C5	C6	C7
0.119556	**0.096373**	0.177238	0.162421	0.170596	0.155586	0.163766
12.4%	**11.2%**	15.1%	14.5%	14.8%	14.2%	14.5%

Adding pairs of Cj*block(replicate) to the IBD response model resulted in the following residual mean squares and coefficients of variation:

	C2	C3	C4	C5	C6	C7
C1	0.078322	0.154536	0.110238	0.135307	0.136667	0.147094
	10.1%	14.1%	11.9%	13.2%	13.3%	13.8%
C2	--	0.120176	**0.068188**	0.130605	0.100498	0.116276
	--	12.5%	**9.4%**	13.0%	11.4%	12.3%
C3	--	--	0.181158	0.237622	0.218425	0.199317
	--	--	15.3%	17.5%	16.8%	16.0%
C4	--	--	--	0.189170	0.197262	0.200890
	--	--	--	15.6%	16.0%	16.1%
C5	--	--	--	--	0.159851	0.159701
	--	--	--	--	14.4%	14.4%
C6	--	--	--	--	--	0.182660
	--	--	--	--	--	15.4%

Adding C2*block(replicate) decreases the residual mean square by 100(1 – 0.096373/0.146648) = 34.3%. Adding the pair C2*block(replicate) and C4*block(replicate) decreased the residual mean square by 100(1 – 0.068188/0.146648) = 53.5% with the C4*block(replicate) term accounting for 100(1 – 0.068188/0.096373) = 29.2% of the decrease.

The response models resulting in minimum residual mean squares for each of the seven characteristics reported in this experiment are given below:

Grainwt = replicate + genotype + block(replicate) + C1*block(replicate) + C7*block(replicate) + error

Rust = replicate + genotype + block(replicate) + C1*block(replicate) + C3*block(replicate) + error

TGW = replicate + genotype + block(replicate) + C2*block(replicate) + C7*block(replicate) + error

grm2 = replicate + genotype + block(replicate) + C1*block(replicate) C7*block(replicate) + error

anthesis = replicate + genotype = block(replicate) + C1*block(replicate) C2*block(replicate) + error

maturity = replicate + genotype + block(replicate) + C3*block(replicate) + C6*block(replicate) + error

CTDX = replicate + genotype + block(replicate) + C2*block(replicate) + C4*block(replicate) + error

The above demonstrates that the response model that best explains the spatial variation present in an experiment must be determined for *each* characteristics measured. One size does *not* fit all! The same response model was obtained for weight of grain and for grain weight per meter squared as it should be since they are essentially the same. It also demonstrates that there is no ordering of polynomial regressions when it comes to explaining spatial variation in experiments. The so-called "hierarchical principle" discussed by Federer (2000) used in this context is misguided and inappropriate.

The data for the above example are available upon request. A SAS GLM code for the above example for the final models is given below. If the blocking variables in the model are considered to be random effects and inter-effect information is to be recovered (and it should be), the code is given for only for the response model for grain weight as it is straight-forward to obtain the code for the remaining characters.

```
roc iml;
  opn8=orpol(1:8,7);
  opn8[,1] = (1:8)`;
  op8=opn8;
  create opn8 from opn8[colname={'COL' 'C1' 'C2' 'C3' 'C4' 'C5'
  'c6' 'C7'}] ;
  append from opn8;
  close opn8; run;
  opn15 = orpol(1:15,14);
  opn15[,1]=(1:15)`;
  op15 = opn15;
  create opn15 from opn15[colname={'ROW' 'R1' 'R2' 'R3' 'R4' 'R5'
  'R6' 'R7' 'R8' 'R9' 'R10' 'R11' 'R12' 'R13' 'R14'}];
  append from opn15;
close opn15;  run;
data reynolds;
  infile 'reynolds.dat';
  input plot block geno rust rep grainwt TGW grm2 anthesis maturity
CTDX  col;
data augbig; set reynolds;
  idx = _n_;
run;
proc sort data = augbig;
  by col; run;
```

```
data augbig;
  merge augbig opn8;
  by col;      run;
proc sort data = augbig;
  by row;  run;
data augbig;
  merge augbig opn15;
  by row;  run;
proc sort data = augbig;
  by idx;  run;
proc glm data= augbig;
   class rep block col geno;
   model grainwt = rep geno block(rep);
   lsmeans geno;
run;
proc glm data= augbig;
   class rep block col geno;
   model grainwt = rep geno block(rep) C1*block(rep) C7*block(rep);
   lsmeans geno;
run;
proc glm data= augbig;
   class rep block col geno;
   model rust = rep geno block(rep) C1*block(rep) C3*block(rep);
   lsmeans geno;
run;
proc glm data= augbig;
   class rep block col geno;
   model TGW = rep geno block(rep) C2*block(rep) C7*block(rep);
   lsmeans geno;
run;
proc glm data= augbig;
   class rep block col geno;
   model grm2 = rep geno block(rep) C1*block(rep) C7*block(rep);
   lsmeans geno;
run;
proc glm data= augbig;
   class rep block col geno;
   model anthesis = rep geno block(rep) C1*block(rep) C2*block(rep);
   lsmeans geno;
run;
proc glm data= augbig;
   class rep block col geno;
   model maturity = rep geno block(rep) C3*block(rep) C6*block(rep);
   lsmeans geno;
run;
proc glm data= augbig;
   class rep block col geno;
   model CTWX = rep geno block(rep) C2*block(rep) C4*block(rep);
   lsmeans geno;
run;
proc mixed data= augbig;
   class rep block col geno;
   model grainwt =  geno block;
   random rep block(rep) C1*block(rep) C7*block(rep);
   lsmeans geno;
run;
```

A 15-ROW BY 12-COLUMN EXPERIMENT DESIGN EXAMPLE

At another site, the 120 wheat genotypes in the preceding example were included in a 15-row by 12-column design along with two check genotypes each replicated 30 times. The 120 genotypes occurred once in the experiment. The experiment design was an augmented row-column experiment design. The polynomial regression method described above is appropriate for this type of layout. For the 60 responses from the two checks, not all row, column, and check effects have non-zero solutions. The rank of the design matrix is two less than required for solution of these effects. It is possible to obtain an analysis of variance, ANOVA, using SAS PROC GLM but not least squares means. Performing this operation resulted in

	Source of variation	Degrees of freedom	Mean square
Type I:	Row	14	35,559
	Column	11	58,885
	Genotype	119	18,630
	Residual	35	14,820
Type III:	Row	12	18,171
	Column	9	18,839
	Genotype	119	18,630

Note that the degrees of freedom for genotype in the Types I and III ANOVAs should have been 121 as there were 122 genotypes in the experiment. Owing to the fact that the rank was two less than needed for solutions for all effects, this shows up in the Type III ANOVA. The coefficient of variation for this analysis is 12.8% and the F-value for genotype to residual mean squares is 1.26.

To obtain genotype means, some functions of row and column effects are required. Use is made of polynomial regression of responses on position, R1 to R12, and of column responses on position, C1 to C10, and interactions of these regression coefficients. Federer, Reynolds, and Crossa (2001) considered only interactions through quartic regressions. The response model given by them is

$$\text{Grain weight} = C1 + C4 + C10 + R2 + C1*R1 + C1*R3 + C2*R2 + C2*R4 + C3*R2 + C3*R4 + C4*R3 + C4*R4 + \text{genotype} + \text{error}$$

This response model resulted in a residual mean square of 6,088 with 46 degrees of freedom. The coefficient of variation is 8.2% and the F-statistic for genotype to residual mean squares is 3.34.

However, if the entire range of row regression by column regression interactions is examined, considerably more of the spatial variability in this experiment can be taken into account. Using the response model

Grain weight = C1 + C4 + C10 + R2 + C1*R1 + C1*R3 + C1*R9 + C2*R2 + C2*R4 + C2*R11 + C3*R2 + C3*R4 + C4*R3 + C3*R4 + C4*R6 + C5*R5 + C5*R10 + C5*R12 + C6*R5 + C6*R7 + C8*R5 + C8+R11 + C9*R12 + C10*R6 + genotype + error,

resulted in a residual mean square of 1,810 with 34 degrees of freedom. Interactions of high degree polynomials are required to account for the spatial variation in this experiment. The coefficient of variation is 4.5% and the F-value for genotype over residual mean squares is 11.5, a considerable change over the previous two response models. The selection of the above response model resulted in 14,820/1,810 = 8.2 times more replication to obtain the same residual mean square than would have been obtained using the textbook row-column-genotype response model. Federer, Reynolds, and Crossa (2001) should have considered higher than fourth degree polynomial regression interactions. The above model resulted in 6,088/1810 = 3.4 times more replication than limiting interactions to fourth degree.

A SAS PROC GLM code for the above response models is given below:

```
proc iml;
  opn12=orpol(1:12,10);
  opn12[,1] = (1:12)`;
  op12=opn12;
  create opn12 from opn12[colname={'COL' 'C1' 'C2' 'C3' 'C4' 'C5'
'C6' 'C7' 'C8' 'C9' 'C10'}]; append from opn12;
  close opn12; run;
  opn15 = orpol(1:15,12) ;
  opn15[,1]=(1:15)`;
  op15 = opn15;
  create opn15 from opn15[colname={'ROW' 'R1' 'R2' 'R3' 'R4' 'R5'
'R6' 'R7' 'R8' 'R9' 'R10' 'R11' 'R12'}]; append from opn15;
close opn15;  run;
data augsite2;
  infile 'c:\my documents\my SAS files\augsite2.dat';
  input  row col genotype grainwt;if (genotype>120)
then new = 0; else new = 1;  if (new) then trtn= 999; else
trtn=genotype;
data augbig; set augsite2;
  idx = _n_;
run;
proc sort data = augbig;
  by col; run;
data augbig;
  merge augbig opn12;
  by col;      run;
proc sort data = augbig;
  by row;  run;
data augbig;
  merge augbig opn15;
  by row;  run;
proc sort data = augbig;
  by idx;  run;
proc glm data = augbig;
   class row col genotype trtn;
   model grainwt =C1 C4 C10 R2 C1*R1 C1*R3 C2*R2 C2*R4 C3*R2 C3*R4
                  C4*R3 C4*R4 genotype;
```

```
run;
proc GLM data = work.augbig ;
  class row col genotype ;
  model grainwt =  C1 C4 C10 R2 C1*R1 C1*R3 C1*R9 C2*R2 C2*R4 C2*R11
            C3*R2 C3*R4 C4*R3 C4*R4 C4*R6 C5*R5 C5*R10 C5*R12 C6*R5
            C6*R7 C8*R5 C8*R11 C9*R12 C10*R6 genotype;
   RUN;

proc glm data = work.augbig;
   class row col genotype;
   model grainwt = row col genotype;
run;
```

A SECOND 15-ROW BY 12-COLUMN EXPERIMENT DESIGN EXANPLE

The 120 wheat genotypes discussed in the above two examples were grown in a 15-row by 12-column experiment design at a third site. The weight of grain for this example has been given in Federer (1998). As in the preceding example the experiment design was such that not all row, column, and genotype effects have solutions as the rank is two less than required. If two row or two column contrasts such as R13 and R14 or C10 and C11 are set equal to zero (eliminated from the model), then solutions for effects may be obtained. However, as pointed out above, the SAS PROC GLM code will do this automatically for an ANOVA. Doing this the residual mean square was 5,630.7 with 35 degrees of freedom and a coefficient of variation of 8.5%. The F-ratio of genotype to residual mean squares was 1.47. Limiting consideration to row-column interaction regressions to fourth degree polynomials as Federer, Crossa, and Franco (1998) did, the residual mean square was 3,449.1 with 44 degrees of freedom and a coefficient of variation of 6.7%. The F-statistic for genotypes was 2.44. Their response model was

$$\text{grainwt} = C1 + C2 + C3 + C4 + C6 + C8 + R1 + R2 + R4 + R10 + C1*R1 + C2*R1 + C3*R1 + \text{genotype} + \text{error},$$

where Cj is the jth polynomial regression coefficient of grain weight on column position and Ri is the ith polynomial regression of grain weight on row position. If interactions of all column regressions by row regressions are screened by the rule used above, the resulting response model equation is

$$\text{grainwt} = C1 + C2 + C3 + C6 + C8 + R1 + R8 + R10 + C1*R1 + C2*R1 + C3*R1 + C2*R5 + C3*R7 + C4*R9 + C5*R10 + C6*R12 + C7*R3 + C7*R11 + C8*R2 + C9*R1 + \text{genotype} + \text{error}$$

Using this response model resulted in a residual mean square of 1,081.4 with 38 degrees of freedom and a coefficient of variation of 3.7%. The genotype F-ratio increased to 8.04. Thus, the preceding response model resulted in a residual mean square with 38 degrees of freedom and 5,630.7/1,081.4 = 5.2 times more replication than the textbook row-column-genotype response model and 3,449.1/1,081.4 = 3.2 times more replication than the Federer, Crossa, and Franco (1998) response model. It is to be noted that the patchier the spatial variation, the higher will be the degree of the polynomial regression interactions required to account for this. The above response model used fewer degrees

of freedom, 20, for blocking variables than did the row-column-genotype response model, 23.

A SAS PROC GLM program for the above example is :

```
proc iml;
  opn12=orpol(1:12,10);
  opn12[,1] = (1:12)`;
  op12=opn12;
  create opn12 from opn12[colname={'COL' 'C1' 'C2' 'C3' 'C4' 'C5'
'C6' 'C7' 'C8' 'C9' 'C0' }]; append from opn12;
  close opn12; run;
  opn15 = orpol(1:15,12) ;
  opn15[,1]=(1:15)`;
  op15 = opn15;
  create opn15 from opn15[colname={'ROW' 'R1' 'R2' 'R3' 'R4' 'R5'
'R6' 'R7' 'R8' 'R9' 'R10' 'R11' 'R12'}]; append from opn15;
close opn15;  run;
data augsite1;
  infile 'c:\my documents\my SAS files\augmen1.sas';
  input site obs col row genotype Y1 Y2 Y3 Y4 Y5 grainwt;if
(genotype>120)then new = 0; else new = 1;  if (new) then trtn= 999;
else trtn=genotype;
data augbig; set augsite1;
  idx = _n_;
run;
proc sort data = augbig;
  by col; run;
data augbig;
  merge augbig opn12;
  by col;      run;
proc sort data = augbig;
  by row;  run;
data augbig;
  merge augbig opn15;
  by row;  run;
proc sort data = augbig;
  by idx;  run;
proc glm data = augbig;
   class row col genotype trtn;
   model grainwt =C1 C2 C3  C6 C8 R1  R8 R10 R1*C1
   R1*C2 R1*C3  C2*R5   C3*R7 C4*R9
   C5*R10  C6*R12  C7*R3 C7*R11
   C8*R2  C9*R1  genotype;
run;
proc glm data = augbig;
   class row col genotype;
   model grainwt = row col genotype;
run;
proc glm data = augbig;
   class row col genotype trtn;
   model grainwt =C1 C2 C3 C4 C6 C8 R1 R2 R4 R8 R10 R1*C1
    R1*C2 R1*C3 genotype;
run;
```

A 8-ROW BY 7-COLUMN EXPERIMENT DESIGN

An experiment described in Federer and Schlottfeldt (1954) was designed as a randomized complete block experiment design (RCBD) with v = 7 treatments and r = 8 complete blocks. However, the experiment was laid out in an 8-row by 7-column arrangement, RCD. Owing to several sandy patches in the experimental area and to unfavorable moisture conditions, there was considerable spatial variation present in the experiment. The RCBD ANOVA resulted in a residual man square of 30,228.2 with 42 degrees of freedom, a coefficient of variation of 17.2%, and an F-value for treatments of 1.51. The RCD ANOVA produced a residual mean square of 7,351.8 with 36 degrees of freedom, a coefficient of variation of 8.5%, and an F-value for treatments of 2.71. The response model

Y = C1 + C2 + C3 + C5 + R1 + R2 + R3 + R5 + R6 + R7 + C1*R1 + C2*R1 + C2*R3 + C3*R4 + C4*R1 + C4*R2 + treatment + error

as used by Federer, Crossa, and Franco (1998) and limiting the investigations to interactions of fourth degree row and column regressions, produced a residual mean square of 4,204.5 with 33 degrees of freedom, a coefficient of variation of 6.4%, and an F-value for treatments of 6.36. Considering interactions of all polynomials resulted in the following response model equation:

Y = C1 + C2 + C3 + C5 + R1 + R2 + R3 + R5 + R6 + R7 + C1*R1 + C2*R1 + C2*R3 + C3*R4 + C4*R1 + C4*R2 + C1*R5 + C3*R5 + C3*R7 + C4*R5 + C4*R7 + C5*R4 + C5*R7 + C6*R2 + C6*R7 + treatment + error.

That is, nine interaction terms were added to the previous model. The resulting residual mean square is 1,320.4 with 24 degrees of freedom. The coefficient of variation is 3.6% and the F-value for treatments is 20.49. As may be observed, tremendous differences exist between the analyses for the different response models. A standard textbook approach would use the RCBD analysis, and consideration of the spatial layout of the experiment would use the RCD analysis. Consideration of differential gradients would result in the response model

Y = row + treatment + C2*row + C3*row + C4*row + error.

The residual mean square for this model was 11,309.9 with 18 degrees of freedom, a coefficient of variation of 10.5%, and an F-value for treatments of 3.78. This model used 31 degrees of freedom for blocking variables, leaving only 18 for the residual. Furthermore, it was not as effective as the row-column-treatment model for controlling spatial variation. Using the next to last model above, was quite effective in accounting for the spatial variation in this experiment and effectively resulted in 30,228.2/1,320.4 = 22.9 times more replication than the RCBD analysis and 7,351.8/1,320.4 = 5.6 times more replication than the RCD analysis.

A SAS PROC GLM code for the above models is given below:

```
options ls = 76;
proc iml;
```

```
   opn4=orpol(1:7,6);
   opn4[,1] = (1:7)`;
   op4= opn4;
   create opn4 from opn4[colname={'COL' 'C1' 'C2' 'C3' 'C4' 'C5'
   'C6'}]; append from opn4;
   close opn4;
run;
   opn3=orpol(1:8,7);
   opn3[,1] = (1:8)`;
   op3 = opn3;
   create opn3 from opn3[colname={'ROW' 'R1' 'R2' 'R3' 'R4' 'R5'
   'R6' 'R7'}]; append from opn3;
   close opn3;
run;
data colrow;
   infile 'c:\my documents\my SAS files\colrow.dat';
   input yield ROW COL T ;
data rcbig; set colrow;
   idx = _n_; run;
proc sort data= rcbig;
   by     COL ; run;
data rcbig;
   merge rcbig opn4;
   by     COL;  run;
proc sort data = rcbig;
   by     ROW;  run;
data rcbig;
   merge rcbig opn3;
   by     ROW;   run;
proc sort data = rcbig; by idx; run;
proc glm data = rcbig;
   class  ROW COL T;
   model yield = row T;

run;
proc glm   data = rcbig;
   class  ROW COL T;
   model yield = row col T;

run;

proc glm data = rcbig;
   class ROW COL T;
  model yield = row T c2*row c3*row c4*row;
run;

proc glm  data = rcbig;
   class  ROW COL T;
   model yield = T c1 c2 c3 c5 r1 r2 r3 r5 r6 r7 c1*r1 c2*r1
  c4*r1 c4*r2 c3*r2 c2*r3 ;
 run;
proc glm data = rcbig;
   class ROW COL T;
   model yield = T c1 c2 c3 c5 r1 r2 r3 r5 r6 r7 c1*r1 c2*r1 c2*r3
   c3*r2 c4*r1 c4*r2 c1*r5  c3*r5
    c3*r7 c4*r5  c4*r7 c5*r4 c5*r7 c6*r2 c6*r7;
run;
```

DISCUSSION

The above examples utilized a fixed effects approach to exploratory model selection. Federer and Wolfinger (2000) have presented two random effects procedures for model selection. A comparison of resulting models using the three procedures with above examples could be made. It is fairly certain that the resulting models would differ. Until the properties of the random effects selection procedures are known, this will not be done. It is known that the Bozivich, Bancroft, and Hartley (1956) procedure has only a small effect on the Type I error. It is possible that F at the 25% level is not optimal for reducing the effect on Type I errors. With present computing power, this needs to be investigated.

The model selection procedure utilized the fixed effects analyses. When obtaining treatment means, one should recover the information from the random blocking effects. As noted from remarks by several anonymous referees, they have a difficult time thinking of regression and gradient coefficients as random blocking effects. They appear to have no difficulty with considering row, column, and block effects as random, but they appear to fail to appreciate the fact that the regressions and gradients are merely functions of row and column effects that are random effects. Hence, in obtaining adjusted treatment means, all blocking effects should be considered as random effects and the information contained in them needs to be recovered in order to utilize efficient procedures.

Different models were obtained for each of the characters analyzed in the first example. This means that an experimenter should perform exploratory model selection for each characteristic being analyzed. Use of computer programs such as those given above, make this is relatively simple matter.

The fixed methods of regression and gradients used in this investigation have known degrees of freedom for the various parameters used in a response model. Several other procedures such as smoothing, Kriging, nearest neighbor, and autoregression have been proposed for spatial analyses (See Federer, Newton, and Altman, 1997, e.g.). The degrees of freedom for each of the parameters used in these methods are usually unknown. Hence, it is difficult to compare their ability to explain spatial variation in comparison to regression and gradient procedures.

Orthogonal polynomial regression was used in the above analyses. This involves using centered values for the independent (covariate) variable, position. It should be noted that interactions of non-centered covariates *will not be* the same as those from centered covariates. Also, instead of using orthogonal polynomial regressions, Fourier regression may be more appropriate in some situations, i.e., when the spatial variation is cyclical in nature.

Some statisticians, e.g., Gilmour (2000), appear to believe that the above exploratory model investigation is "post blocking that has gone too far." If the variation can be accounted for and if there are sufficient degrees of freedom, say 20-30, associated with the residual mean square, there should be no reason why a data analyst should not use procedures such as those described herein. Others believe that regression coefficients and gradients should always be considered as fixed effects. They appear to fail to realize that as used herein, they are functions of random variable effects and hence should be

considered to be random effects. In field layouts, there are no valid reasons to consider that there will be a single regression and that all variation follows an orderly and systematical pattern. Even though an experimenter may try to select a uniform area in which to conduct the experiment, this is not always possible, e.g., conducting experiments on farmer's fields.

LITERATURE CITED

Bozivich, H., T. A. Bancroft, and H. O. Hartley (1956). Power of analysis of variance test procedures for certain incompletely specified models. Annals of Mathematical Statistics 27:1017-1043.

Federer, W. T. (1998). Recovery of interblock, intergradient, and intervariety information in incomplete block and lattice rectangle designed experiments. Biometrics 54:471-481.

Federer, W. T. (2000). Some misconceptions about SAS PROC GLM and PROC MIXED outputs, model selection, and trend analyses (with discussion). BU-1475-M in the Technical Report Series of the Department of Biometrics, Cornell University, Ithaca, New York 14853.

Federer, W. T., J. Crossa, and J. Franco (1998). Forms of spatial analyses with mixed model effects and exploratory model selection. BU-1406-M in the Technical Report Series of the Department of Biometrics, Cornell University, Ithaca, New York 14853.

Federer, W. T., E. A. Newton, and N. S. Altman (1997). Combining standard block analyses with spatial analyses under a random effects model. In *Modelling Longitudinal and Spatially Correlated Data Methods: Applications and Future Directions* (T. G. Gregoire *et al.*, editors) Springer, New York, Heidelburg, pp. 373-386.

Federer, W. T., M. Reynolds, and J. Crossa (2001). Combining results from augmented designs over sites. Agronomy Journal 93:389-395.

Federer, W. T. and R. D. Wolfinger (2000). Methods of parameter selection in exploratory model analyses. BU-1537-M in the Technical Report Series of the Department of Biometrics, Cornell University, Ithaca, New York 14753.

Gilmour, A. R. (2000). Post blocking has gone too far! Recovery of information and spatial analyses for field experiments. Biometrics 56:944-946